

Academic team formation as evolving hypergraphs

Carla Taramasco^{*,†,‡}

Jean-Philippe Cointet^{*,†,§}

Camille Roth^{*,†,¶,||}

[PREPRINT — PAPER TO APPEAR IN **SCIENTOMETRICS**]

Abstract

This paper quantitatively explores the social and socio-semantic patterns of constitution of academic collaboration teams. To this end, we broadly underline two critical features of social networks of knowledge-based collaboration: first, they essentially consist of group-level interactions which call for team-centered approaches. Formally, this induces the use of *hypergraphs* and *n*-adic interactions, rather than traditional dyadic frameworks of interaction such as *graphs*, binding only pairs of agents. Second, we advocate the joint consideration of structural and semantic features, as collaborations are allegedly constrained by both of them. Considering these provisions, we propose a framework which principally enables us to empirically test a series of hypotheses related to academic team formation patterns. In particular, we exhibit and characterize the influence of an implicit group structure driving recurrent team formation processes. On the whole, innovative production does not appear to be correlated with more original teams, while a polarization appears between groups composed of experts only or non-experts only, altogether corresponding to collectives with a high rate of repeated interactions.

1 Introduction

The mechanisms of academic collaboration are the focus of a long and established tradition of research

^{*}ISC-PIF (Institut des Systèmes Complexes – Paris-Île-de-France). 56, rue Lhomond, 75005 Paris, France.

[†]CREA (CNRS/Ecole Polytechnique, France). CREA/ENSTA, 45 Bd Victor, 75015 Paris, France

[‡]DECOM (Universidad de Valparaíso, Chile). Avenida Gran Bretaña, 1091 Playa Ancha. Valparaíso, Chile

[§]INRA SenS (Sciences en Société) - IFRIS. 5, Bd Descartes. 77420 Champs-sur-Marne, France

[¶]CAMS (CNRS/EHESS, France). EHESS/CNRS, 54 Bd Raspail, 75006 Paris, France

^{||}CRESS (U. Surrey, GB). University of Surrey, Guildford GU2 7XH, United Kingdom

Emails:

roth@ehess.fr (*corresponding author*)

Carla.Taramasco@polytechnique.edu

Jean-Philippe.Cointet@polytechnique.edu

(Katz & Martin, 1997), from qualitative studies on cooperation and co-optation behaviors (Crane, 1969; Chubin, 1976; Latour & Woolgar, 1979) to more quantitative approaches (deB. Beaver & Rosen, 1978–1979; deB. Beaver, 1986; Melin & Persson, 1996). The latter includes network-based studies, which are generally aiming at understanding the structural determinants and patterns of collaboration (Mullins, 1972; Newman, 2001; Barabási et al., 2002; Moody, 2004; Wagner & Leydesdorff, 2005; Leahey & Reikowsky, 2008). In this case, the quantitative formal framework of choice is the social network of dyadic interactions, addressing questions related to how ego-centered characteristics, *in the broad sense*, influence the likelihood of being involved in a collaboration.

The Team Level and Networks

Network studies, specifically in the context of scientific collaboration, indeed often focus on the level of the individual in spite of a large amount of work on the question of group cohesiveness (Lott & Lott, 1965; Bollen & Hoyle, 1990; Friedkin, 2004). There are wider implications of this focus on the ego-centered level:

- By aiming at describing individual behavioral patterns, this perspective may overlook the influence of characteristics expressable at the meso-level of the team itself. In particular, by focusing on dyadic interactions and relational patterns between ego and alter(s), the presence of ego in a given collaboration is interpreted as a function of the characteristics of ego and those of alter(s), and of the characteristics of the various dyads between ego and alter(s).
- Further, the creation of a group results from a complex agreement and arrangement between all its members, who jointly decide to collaborate. As such, even when assuming that the behavior of ego may depend on non-dyadic, team-level characteristics, interpreting team formation processes as a sum of individual rationalities may oftentimes seem difficult, or irrelevant. Put differently, there are regularities in team formation processes which are difficult to ascribe specifically back to individuals; it may appear more natural

and consistent to appraise the underpinnings of group formation at the group level.¹

To sum up, when dyadic frameworks are involved, collaboration teams are appraised under the lens of multiple one-to-one interactions. It should be no surprise: social network literature is itself overwhelmingly concerned with dyadic links. However, a sizeable portion of sociology, starting with Simmel (1898), has long been concerned by wider frameworks of interactions, or so-called “social circles”, which some authors have formalized to take directly into account non-dyadic relationships: Breiger (1974, 1990), for instance, proposed to use bipartite graphs to represent and analyze ties between actors and social groups. Focusing on the group-level, Ruef (2002) quantitatively examined the contribution of several factors including gender, status, or ethnicity, in the preferential constitution of business founding teams. In a review study, Freeman (2003) explored various approaches previously adopted in mathematical sociology to model two-mode data in order to account for the presence of subsets of people participating altogether in (subsets of) identical events.

In this respect, it therefore first appears that academic collaboration choices and dynamics should be characterized by investigating the meso-level of team formation. More precisely, it should be fruitful to focus on *teams* rather than pairs of agents interacting together, thus advocating the use of *hypergraphs* or bipartite graphs rather than traditional frameworks based on graphs. Hypergraphs indeed feature *hyperlinks* which connect arbitrary numbers of agents, while graphs feature *links* which connect only *pairs* of agents. In other words, considering hypergraphs prevents making the superfluous and plausibly debatable assumption that teams are equivalent to complete subgraphs featuring one-to-one interactions between all its members (i.e. assuming for instance that a triad is equivalent to three dyads).

Hybrid Networks of Actors and Concepts

Secondly, collaboration massively depends on cognitive properties, in particular some cognitive fit between team members, as agents plausibly compose teams in order to gather complementary competences. For instance, some economic models of knowledge creation consider matching rules based on the similarity of agent profiles, as elements of a vector space, to explain economic network structure (Cowan et al., 2002). In other words, equal attention should be given

¹Note that what we call a “team” here actually relates to a group that is involved in the production of an academic paper, i.e. the team of coauthors that produces it; it does not correspond to the more or less explicit notion of team that may exist in some research labs.

to social and semantic features, which are traditionally left apart in the literature, although the existence of homophily-driven interactions has been underlined in numerous works (McPherson et al., 2001).

Our main hypothesis is that one cannot correctly understand the underlying social processes if both social and semantic dimensions of, e.g., scientific activity, are not considered as two interdependent dynamics (Roth, 2006; Roth & Cointet, 2010). Going further, we construe scientific dynamics as made of groupings of both agents and concepts: the *epistemic* dynamics, i.e. the collective scientific knowledge construction, is made of events which simultaneously involve compounds of actors and concepts. In line with the program introduced by Callon (1986), we will appraise scientific dynamics as made of constant reconfiguration and re-negotiation of collectives of both humans and non-humans.

In this respect and more broadly, in addition to focusing on teams, we thus advocate the enrichment of the notion of team by *considering teams as joint groupings of both agents and semantic items*.

Knowledge-based teamwork

The interest in the social epistemology of academic communities also has a broader reach. As a knowledge production arena, science is indeed likely to share features found in other collaborative knowledge creation contexts.

(i) *Collaboration in knowledge production systems.*

This issue may shed light, to some extent, on the interaction processes underlying, broadly, collaborative knowledge production. These contexts indeed define a particularly common class of social networks of collaboration, where agents jointly and collectively interact for purposes of knowledge production, in the broad sense. This encompasses activist groups and political epistemic communities (Ruggie, 1975; Haas, 1992), scientific communities (deB. Beaver & Rosen, 1978–1979; Laband & Tollison, 2000; Jones et al., 2008; Stokols et al., 2008; Leahy & Reikowsky, 2008) and more specifically research projects (Larédo, 1995, 1998), open-source development communities (Kogut & Metiu, 2001) and discussion lists and forums (Constant et al., 1996; Welser et al., 2007), wiki platform-mediated communities (Bryant et al., 2005; Levrel, 2006), artists gathering for a theater performance (Uzzi & Spiro, 2005) or making a movie (Faulkner & Anderson, 1987; Ramasco et al., 2004), board members making collective decisions (Davis & Greve, 1996).

(ii) *Collaboration in teams.*

This kind of relatively autonomous collaboration mode has to be understood in a context where traditionally vertical and hierarchical organizations have recently been functioning in increasingly horizontal and networked ways (Powell, 1990; Miles & Snow, 1996; Smangs, 2006). This contemporary so-called “network governance” involves dynamic coalitions of actors both at organizational and individual levels, increase of teamwork and frequent group reconfigurations (Jones et al., 1997). This shift is particularly sensible in contexts where agents are relatively free to group to form casual alliances and where collaboration sometimes appears to be self-organized.

In this respect, science appears to be a prototypical case of such teamwork-based systems (deB. Beaver, 1986; Adams et al., 2005; Wuchty et al., 2007) — scientific knowledge production essentially involves events where researchers jointly work to manipulate and introduce concepts. It is additionally one of the most accomplished context of *knowledge-based* collaboration, as well as one of the most explicit, by its very stigmergic² nature: papers indeed constitute a concrete, often public instance of these gatherings and therefore provide an opportunity to understand the impact of these collaborations on the dynamics of science. On the empirical side, we thus rely on large bibliographic databases.

As such, our approach does not pretend to embrace the whole complexity of knowledge-intensive organizations, in particular the intricate co-evolutionary processes existing between formal organizations and more local team-based and individual-based decisions (Lazega et al., 2008). However, the methodology we propose is able to shed some original light on portions of the dynamics of these knowledge production systems.

The paper is organized as follows: in Sec. 2, we present the framework and support several hypotheses on socio-semantic team-based collaboration, Sec. 3 introduces the protocol and methods, while Sec. 4 presents the results, which we then discuss in light of the initially proposed hypotheses.

2 Framework

As follows from the introduction, we hence argue that two features are key in extending the understanding of, one hand, collaboration networks, and on the

²“Stigmergic”: that is, leaving *traces* susceptible to guide the work of others. For an extensive discussion of this notion, see Karsai & Penzes (1993).

other hand and additionally, knowledge production networks:

1. Group effects underlie and partially determine dyadic interactions: affiliation to teams of collaboration, membership in identical epistemic communities, for instance, structure and influence the very formation of these interactions.
2. In the case of social networks of knowledge, these underlying groups are both social (work communities) and semantic (epistemic communities). In particular, the choice of collaboration partners is likely to highly depend on cognitive similarity.

More to the point, in terms of strictly social and strictly semantic associations, we first aim at checking the following simple hypotheses, by comparing what happens empirically with what would have happened if teams had been formed strictly by chance (i.e. by comparing empirical teams with a null-model featuring random compositions of teams).

- (H1). Teams with a high rate of interaction repetition should be more likely, as could be expected because of social cohesion (Bollen & Hoyle, 1990; McPherson & Smith-Lovin, 2002; Friedkin, 2004) or organizational constraints (Rodriguez & Pepe, 2008).
- (H2). Teams where a high proportion of concepts are repeatedly associated should be more likely — as assumed by co-word analysis (Callon et al., 1986; Noyons & van Raan, 1998), where frequent associations of terms are supposed to define conceptual cores and field boundaries.
- (H3). Papers with a higher semantic originality (i.e. new association of concepts) should be those where there is a higher number of new interactions.³ Put differently, as suggested by social and semantic repetitions assumed by H1 and H2, teams with a high number of repeated interactions should tend to produce papers that have smaller semantic/topical originality; which in some sense belong to a narrower subfield of research (Leahey & Reikowsky, 2008).

Then, we appraise the socio-semantic composition of teams. We more precisely focus on the distinction

³As Callon (1994, p.414) sums up from the existing literature,

“The more numerous and different these heterogeneous collectives are, the more the reconfigurations produced are themselves varied”

between agents who are already familiar with some concepts involved in the interaction, and those who are not. This approach will more broadly inform us about the cognitive specialization of teams.

(HI). Because of both scientific specialization (Chubin, 1976) and homophily (McPherson et al., 2001; Stokols et al., 2008), teams gathering around a given topic should generally involve more individuals knowledgeable about this given topic.

(HII). Teams with a balanced composition of experts in a given field should produce more innovation (Ancona & Caldwell, 1992), which in terms of networks could be translated into:

- more semantic originality, i.e. novel associations of concepts,
- more social originality, i.e. novel interactions between agents.

3 Protocol and methods

In line with this focus on socio-semantic aspects, we will thus endeavor at exhibiting how new teams are formed by considering both social and conceptual past acquaintances of scientists involved in new collaborations. We will concretely describe the semantic dimension in terms of attributes qualifying topics of interest of authors and the social dimension as structural and relational properties in the dynamic collaboration network — which altogether will enable us to confirm or refute the previous set of hypotheses.

3.1 Datasets

Our empirical analysis focuses on collaboration databases, which reveal a large part of the underlying collaboration activity, including social links between individuals or conceptual acquaintances of each individual (i.e. details regarding which topics which agents are familiar with). These datasets provide temporal information on teams, gathering agents and the topics they work on, assuming that topics are described by the very terms used in paper abstract. For each dataset, we focus on a set of no more than a hundred of relevant terms. These terms are selected with the help of an expert of the corresponding field and are such that they appropriately cover the most significant topics of each field.

We use the following datasets, defined either from a semantic perspective (using e.g. field names) or from a social perspective (using e.g. scientific assemblies), and involving both large and small communities:

1. Embryologists working within a given and well-determined subfield — the zebrafish, on a period of 20 years (1985–2004). Data was extracted from the publicly available database *Medline*, which eventually yields a dataset of 6,145 articles (13 084 authors, 71 word classes).
2. Scientists working on rabies from the same kind of *MedLine* extraction as for zebrafish embryologists — the observed period spans from 1985 to 2007. This ends up with 4 648 events (9 684 authors, 70 word classes).
3. Scientific committee members for JEMRA meetings⁴: this dataset includes the publications of an initial set of 168 scientists involved in these meetings, gathered from 1985 to 2007. This ends up with 5 893 papers (15 375 authors, 69 word classes).
4. Scientific committees members for JECFA meetings⁵: similarly, publications of an initial set of 178 scientists are gathered from 1985 to 2007. This ends up with 8 685 papers (21 195 authors, 85 word classes).

3.2 Hypergraph-based definitions

Now, these agents and concepts *formally* define an evolving hypergraph where each article is a hybrid hyperlink gathering both authors and the topics involved in the collaboration, as partly exemplified by Fig. ??.

In what follows, we describe comprehensively our formal framework (Sec. 3.2.1), which, basically, allows us to gather both agents and concepts in a dynamic setting and to define which agents are new, or not (*newcomers* vs. *veterans*), which concepts are new, or not (*novelties* vs. *standards*), and which agents have used which concepts in the past, or not (*neophyte* or *experts*).

Building upon these definitions, we will then propose a series of *hypergraphic* measures (Sec. 3.2.2) — that is, measures at the level of teams, or non-dyadic measures — which cover the proportion of experts in a given collaboration (*expertise ratio*) and the originality of participants in a team (*hypergraphic repetition*, i.e. describing to what extent a team does gather agents, or concepts, which were jointly associated, at the team-level, in previous periods). For instance, a team with an expertise ratio of one will be such that all agents are experts; a team with a hypergraphic repetition of one, in terms of agents, will be such that all

⁴Joint FAO/WHO Expert Meetings on Microbiological Risk Assessment, http://www.fao.org/ag/agn/agns/jemra_index_en.asp

⁵Joint FAO/WHO Expert Committee on Food Additives, <http://www.who.int/ipcs/food/jecfa>

its agents will have *altogether* previously collaborated (it is zero in case none of the agents have previously been associated).

Then, we present a methodology (Sec. 3.2.3) for computing how much the empirical data diverges from a random setting with a comparison between the actual observed data and a uniform *null-model of hypergraph evolution*. Put simply, we will appraise how much teams with, *e.g.*, a given hypergraphic repetition ratio, are forming significantly more often than could be expected by chance. This latter tool will be the cornerstone of the empirical testing of hypotheses 1-2-3 & I-II.

3.2.1 Objects

Hypergraphs.

Formally, a *hypergraph* features *nodes* and *hyperlinks*, which describe n -adic interactions among any subset of nodes. It is therefore a generalization of the notion of graph whose links only describe dyadic interactions, *i.e.* between pairs of nodes. As such, any hyperlink corresponds to any grouping of agents and any kind of social circle: it may describe social events, organizations, families, teams, etc. A hypergraph is also isomorphic to a bipartite graph, where agents on one side are connected to various affiliations, groups or events on the other side; as such a structure which reifies the duality of social groups (Breiger, 1974; Freeman, 2003). See Fig. ??.

Beyond the simple observation of the structure of such networks, several studies have endeavored at reconstructing structural properties typically induced by the hypergraphic setting — namely, that agents interact within groups of some sort — rather than using dyadic interactions only: in this direction Newman et al. (2001); Ramasco et al. (2004); Guimera et al. (2005), *inter alia*, examine the structure of a social network whose dyadic links stem from teams — team composition is first empirically appraised then stylized and used as a basis for what essentially is a clique addition process. In these models however the focus remains on dyadic relationships or dyadic interaction behaviors, rather than truly hypergraphic measures.

In contrast, the focal level of analysis of the present study is the hypergraph and its hyperlinks.

Epistemic hypergraphs.

To bind the social and semantic aspects, we introduce the notion of *epistemic hypergraph* \mathfrak{E}_t using:

- (i) a set of agents \mathcal{A} ,
- (ii) a set of concepts \mathcal{C} , and

- (iii) the epistemic hypergraph itself $\mathfrak{E} \subseteq \mathcal{P}(\mathcal{A} \cup \mathcal{C})$, describing the joint appearance of agents and concepts, and henceforth the usage of the latter by the former, where each collaboration is a hyperlink $\mathfrak{e} \in \mathcal{P}(\mathcal{A} \cup \mathcal{C})$.

As such, an “*epistemic hypergraph*” is properly defined by a triple $(\mathcal{A}, \mathcal{C}, \mathfrak{E})$. Dynamic epistemic hypergraphs are indexed with time, \mathfrak{E}_t , and are considered to be growing: $t < t' \Rightarrow \mathfrak{E}_t \subseteq \mathfrak{E}_{t'}$.

At each timestep, new teams are formed and thus hyperlinks appear, we denote this set by $\Delta\mathfrak{E}_t$, such that $\mathfrak{E}_t = \mathfrak{E}_{t-1} \cup \Delta\mathfrak{E}_t$. Note that $\Delta\mathfrak{E}_t$ is not necessarily equal to $\mathfrak{E}_t \setminus \mathfrak{E}_{t-1}$ since some teams forming at t may already have appeared in \mathfrak{E}_{t-1} .

See an illustration of this framework on Fig. ??.

We also define a projection operation for hyperlinks: given a hyperlink $\mathfrak{e} \in \mathfrak{E}_t$ and a subset $E \subseteq \mathcal{A} \cup \mathcal{C}$, the projection of \mathfrak{e} on the set E is noted $\mathfrak{e}^E = \mathfrak{e} \cap E$. For instance, the fact that all hyperlinks contain at least one agent translates as $\forall \mathfrak{e}, \mathfrak{e}^{\mathcal{A}} \neq \emptyset$.

We can thus define a (dynamic) collaboration hypergraph $\{\mathfrak{e}^{\mathcal{A}} \mid \mathfrak{e} \in \mathfrak{E}_t\} = \mathfrak{A}_t \subseteq \mathcal{P}(\mathcal{A})$ whose hyperlinks connect team members, and a semantic hypergraph $\{\mathfrak{e}^{\mathcal{C}} \mid \mathfrak{e} \in \mathfrak{E}_t\} = \mathfrak{C}_t \subseteq \mathcal{P}(\mathcal{C})$ whose hyperlinks are sets of concepts mentioned in a given collaboration. In particular, \mathfrak{A}_t is isomorphic to a bipartite graph of collaboration, traditional in the literature (Newman et al., 2001; Guimera et al., 2005).

Neophytes and newcomers.

We say that an agent a is, at t , a “*neophyte in a given concept* $c \in \mathcal{C}$ ” if s/he has never used c at t : formally, if $\nexists \mathfrak{e} \in \mathfrak{E}_{t-1}, \{a, c\} \subseteq \mathfrak{e}$. Otherwise, s/he is called an “*expert*”.

We say that an agent a is a “*newcomer*” if s/he has never published before t , which is equivalent to say that $\nexists \mathfrak{e} \in \mathfrak{E}_{t-1}, a \in \mathfrak{e}$. Otherwise, s/he is called a “*veteran*”.

Similarly, we say that a concept c is a “*novelty*” at t if all agents are neophyte in this concept: $\nexists \mathfrak{e} \in \mathfrak{E}_{t-1}, c \in \mathfrak{e}$. Otherwise, it is a “*standard*”.

3.2.2 Measures

Homogeneity of teams and expertise ratio.

Given these basic concepts, we may first examine the composition of teams using a simple hypergraphic measure pertaining to the composition of teams in terms of a simple proportion of experts: “how much are teams made of people familiar or not with a given concept which is used by the team?”.

We call this proportion *expertise ratio*, noted “ ξ ”; for example, a paper on “ants” where half of the authors already worked on ants has a ratio of expertise

in “ants” of .5. Formally, the expertise ratio $\xi_{c,t}(\mathbf{e})$ in concept $c \in \mathbf{e}^C$ at time t of team \mathbf{e} is given by:

$$\xi_{c,t}(\mathbf{e}) = \frac{|\{a \in \mathbf{e}^A \mid a \text{ is an expert in } c\}|}{|\{a \in \mathbf{e}^A\}|}$$

This notion, derived from the composition of a given team in terms of experts vs. neophytes in a given concept, expresses the socio-conceptual homogeneity of a team. See Fig. ??.

Hypergraphic repetition.

We may also express the degree of originality of the composition of a team and its subsequent groupings by measuring, in the broad sense, the proportion of already-existing associations of items, be it agents or concepts. More to the point, we may talk of *social originality* by describing the rate of new associations of agents in a given team; or, dually, we will denote *conceptual originality* by describing the proportion of new associations of concepts in a paper.⁶

More precisely, in the dyadic case, an interaction is said to be repeated if the two nodes already jointly appeared in a previous collaboration. We extend this notion to the hypergraphic case:

- We first say that a set of nodes has “*previously co-occurred*” if there is at least one previously-existing ($< t$) hyperlink including this set. We define the corresponding function ρ_t as follows:

$$\rho_t(\mathbf{e}) = \begin{cases} 1 & \text{if } \exists \mathbf{e}' \in \mathfrak{E}_{t-1}, \mathbf{e} \subseteq \mathbf{e}' \\ 0 & \text{otherwise.} \end{cases}$$

Thus, for instance, if a and a' never collaborated at t , we have $\rho_t(\{a, a'\}) = 0$.

- The notion of hypergraphic repetition is properly defined for veteran agents and/or standard concepts — by definition, repetition cannot occur with newcomers or novelties.

Therefore, in the following formulas, hyperlinks \mathbf{e} must be such that $\forall e \in \mathbf{e}, \exists \mathbf{e}' \in \mathfrak{E}_{t-1}$ such that $e \in \mathbf{e}'$. In other words, we ensure the use of such hyperlinks by considering, $\forall \mathbf{e} \in \mathfrak{E}_t$, truncated hyperlinks $\underline{\mathbf{e}}$ restrained to the set of previously-existing nodes, i.e.:

$$\underline{\mathbf{e}} = \mathbf{e} \cap \bigcup_{\mathbf{e}' \in \mathfrak{E}_{t-1}} \mathbf{e}'$$

We then compute the *hypergraphic rate of repetition for a hyperlink* $\mathbf{e} \in \mathfrak{E}_t$ as the proportion

⁶In which case, new concept associations are *new* with respect to the whole system, consistently with the social case: i.e. this refers to concept associations which never existed in any paper of the preceding periods.

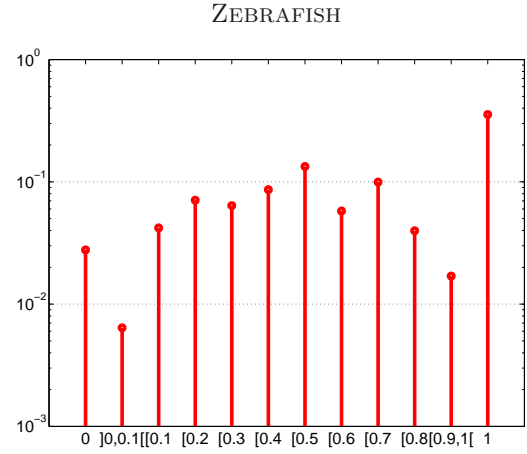


Figure 1: Empirical distribution of the hypergraphic repetition rate for **concepts**, $r_t(\mathbf{e}^C)$.

of subsets of this hyperlink that have previously co-occurred:

$$\begin{aligned} r_t(\mathbf{e}) &= \frac{1}{2^{|\underline{\mathbf{e}}|} - |\underline{\mathbf{e}}| - 1} \sum_{\substack{\mathbf{e}' \subseteq \underline{\mathbf{e}} \\ |\mathbf{e}'| \geq 2}} \rho_t(\mathbf{e}') \\ &= r_t(\underline{\mathbf{e}}) \end{aligned}$$

Depending on the objectives, it might be appropriate to weight the relative importance of each subset of hyperlink \mathbf{e} in the sums, for instance according to their size: for a discussion on weighting functions, see Appendix A.

Let us consider the following example: given a new collaboration \mathbf{e} forming at t , $r_t(\mathbf{e}^C)$ thus measures its hypergraphic concept repetition, i.e. how much the concepts of \mathbf{e}^C have been jointly associated, altogether, in previous periods. Eventually, we may plot the distribution of such values r_t for all teams, as shown in Fig. 1. Put simply, it shows that about a third of teams have a hypergraphic conceptual repetition of 1, i.e. all their concepts \mathbf{e}^C have already *jointly* been used in the past.

3.2.3 Estimating propensities of team formation

Null-model of hypergraph.

A null-model of new teams based on agents (resp. concepts) is defined such that, at each period t , we randomly create new teams respecting empirically-observed numbers of agents (resp. concepts) *and* their respective numbers of team participations. What is fundamentally randomized is the exact composition of teams in terms of who is collaborating with whom: in

our null-model, *team members are basically reshuffled*. Put differently, the null-model expresses the composition of teams as would be happening *by chance*.

In other words and more practically,

- we empirically measure:
 1. the size of new teams appearing at t , i.e. the distribution of $|\mathbf{e}^{\mathcal{A}}|$ (resp. $|\mathbf{e}^{\mathcal{C}}|$) for $\mathbf{e} \in \Delta\mathfrak{E}_t$,
 2. for every element $e \in \mathcal{A}$ (resp. $e \in \mathcal{C}$), the number of times it appears in newly-formed teams, i.e.:

$$|\{\mathbf{e} \in \Delta\mathfrak{E}_t \text{ such that } \mathbf{e} \ni e\}|$$

- we then generate an artificial, uniformly random set of new teams $\widetilde{\Delta\mathfrak{E}_t} \subset \mathcal{P}(\mathcal{A} \cup \mathcal{C})$ which respects above-mentioned distributions, that is:
 1. same distribution of sizes of new hyperlinks,
 2. same distribution of participations of elements in these new hyperlinks.

In the remainder, we examine and compare the properties of the empirical $\Delta\mathfrak{E}_t$ and the randomly-created $\widetilde{\Delta\mathfrak{E}_t}$.

Propensity.

In particular, we define the propensity of team formation with respect to a given function f of a hyperlink (e.g. the hypergraphic rate of repetition) as, for each possible value x of the function, the ratio between the observed number of new hyperlinks (events) \mathbf{e} such that $f(\mathbf{e}) = x$ and the randomly-created number of such events:

$$\Pi_t(x) = \frac{|\{\mathbf{e} \in \Delta\mathfrak{E}_t \text{ such that } f(\mathbf{e}) = x\}|}{|\{\mathbf{e} \in \widetilde{\Delta\mathfrak{E}_t} \text{ such that } f(\mathbf{e}) = x\}|} \quad (1)$$

Obviously, if this quantity is above 1 for a certain value of x , we say that this type of team empirically occurs more than expected; otherwise, less.

4 Results

We may now empirically appraise hypotheses 1-2-3 & I-II.

4.1 Simulation of the null-model

We start by measuring the propensity of team formation, first with respect to simple expertise ratios and, second, with respect to hypergraphic repetition rates. To this end, we simulate 2,500 instances

of above-defined null-model-based epistemic hypergraphs, which are therefore random hypergraphs.⁷ We then compare the composition of teams thus obtained with that of the empirical data.

Expertise ratio: socio-semantic homogeneity/heterogeneity

Distinguishing agents who have already been associated with a concept (“experts”) and agents who are not yet associated (“neophytes”), we thus assess whether real teams involve agents of mixed backgrounds or not, relatively to a randomly-built set of teams. Details of this comparison are displayed on Fig. 2 for the *zebrafish* case, which illustrates the composition of teams for various levels of expertise ratios, in both the real and random cases. Corresponding propensities, for both cases, are shown on Fig. 3: their shapes are consistent across all datasets and consist of a U-shaped curve above 1 for extreme values of expertise ratios (towards 0 and towards 1) and below 1 for central values (typically, from 0.1 to ca. 0.4 – 0.5).

Empirically, we thus observe that there is a significantly high propensity of formation of teams composed of either experts only or newcomers only, with a significantly lower propensity for mixed teams. Teams involving a mixed proportion of experts and newcomers are thus less frequent than they should be.

Hypergraphic rate of repetition: social or semantic homogeneity/heterogeneity

Measuring now propensities of group formation with respect to hypergraphic rates of repetition, we can empirically exhibit the existence and influence of an implicit group structure which drives recurrent team formation — this group structure exists along the two above-mentioned dimensions:

- *Social homogeneity/heterogeneity*: With respect to agents, the hypergraphic rate of repetition measures the extent to which a team features repeated interactions among former collaborators. Once again, our results have to be compared to the null hypothesis for which teams are formed randomly. Figure 4–*top* features the corresponding propensities which are several orders of magnitude higher than 1 for teams with a non-negligible proportion of such repetitions ($r > .1$)
- *Conceptual homogeneity/heterogeneity*: Similarly, we measure the propensity of team formation with respect to repeated concept associa-

⁷For reasons of computational complexity, we consider event sizes not greater than 10 agents and 10 concepts — with this constraint we still consider no less than 89% of the total original number of teams.

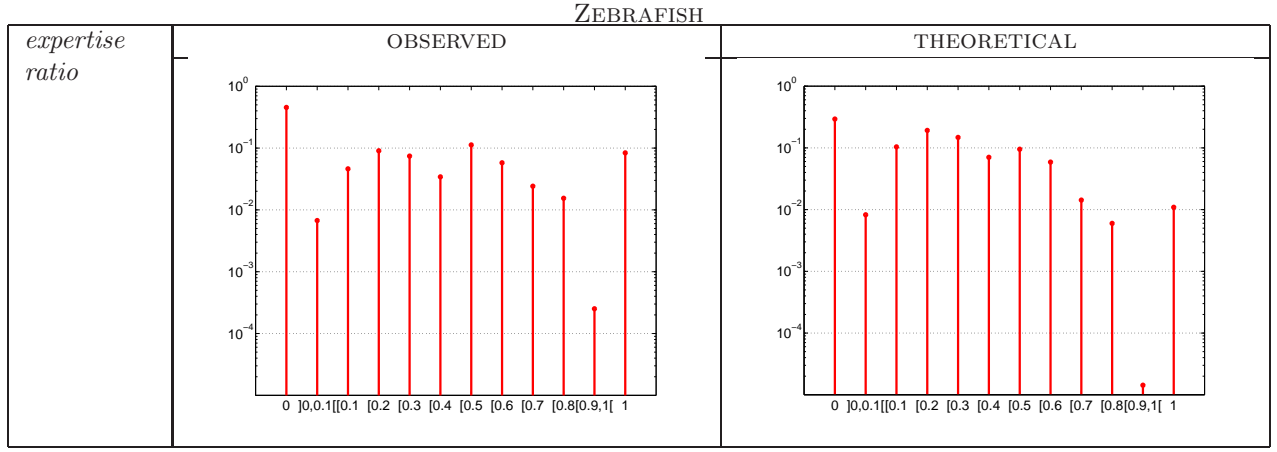


Figure 2: Probability distribution of the expertise ratio on all teams aggregated over all years and all concepts (*left*: observed, *right*: theoretical). The computation of propensities below will be based on the ratio of such observed distributions over theoretical ones.

tions, addressing the following issue: “are there cores of concepts which are likely to be recurrently associated, given that they were previously jointly used in previous papers?” Results, shown on Fig. 4–*bottom*, demonstrate again (and even in a stronger fashion than in the social case) that there is a significant bias towards gathering *groups of concepts* which were previously associated.

4.2 Discussion of hypotheses

It is now possible to review and check the aforementioned hypotheses. As follows from Fig. 4, it is clear that **(H1)** and **(H2)** are *quantitatively* confirmed: teams with a high proportion of interaction repetitions or with a high proportion of repeated conceptual associations are much more likely than should be expected by chance.

Additionally, and irrespective of the simulation model, we check if there is a correlation between semantic and social hypergraphic rates of repetition. As shown on Fig. 5, there seems to be no correlation between social and semantic originality in a collaboration (in our datasets, which come from varied backgrounds but are also focused on particular epistemic communities). This invalidates **(H3)**: in other words, contrarily to intuition, new semantic associations do not stem more from original teams than from repeated teams. In other words, semantic innovation is as likely from agents who, globally, previously collaborated, as from new collaborations.⁸

⁸This does not mean, however, that the backgrounds of previous collaborators who are causing semantic innovation should necessarily be similar (semantic innovation might indeed come from repeated collaboration with individuals who have varied

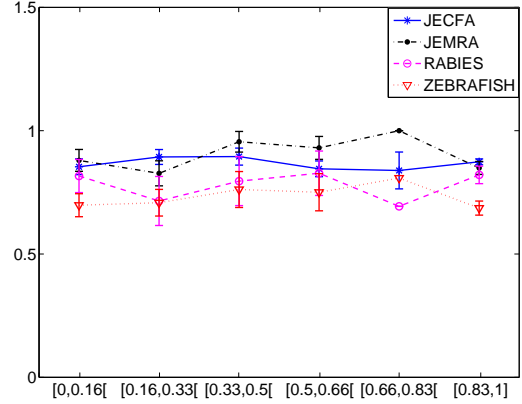


Figure 5: Average semantic hypergraphic repetition ratio (*y-axis*) for a given range of social hypergraphic repetition ratio (*x-axis*). (Error bars correspond to 95% confidence intervals with respect to averages on each repetition ratio bin (in abscissa), such as e.g. $[0, 0.1[$.)

As regards expertise, **(HI)** — “teams gathering around a given topic should involve more individuals knowledgeable about it” — is partially confirmed and partially contradicted by the empirical evidence. Firstly, teams with a high proportion of experts in a concept involved in the collaboration are much more likely, as shown on the right side of each graph on Fig. 3, whose values are significantly above 1.

Yet, secondly, teams with a very small proportion of experts regarding a concept, i.e. high proportion semantic backgrounds).

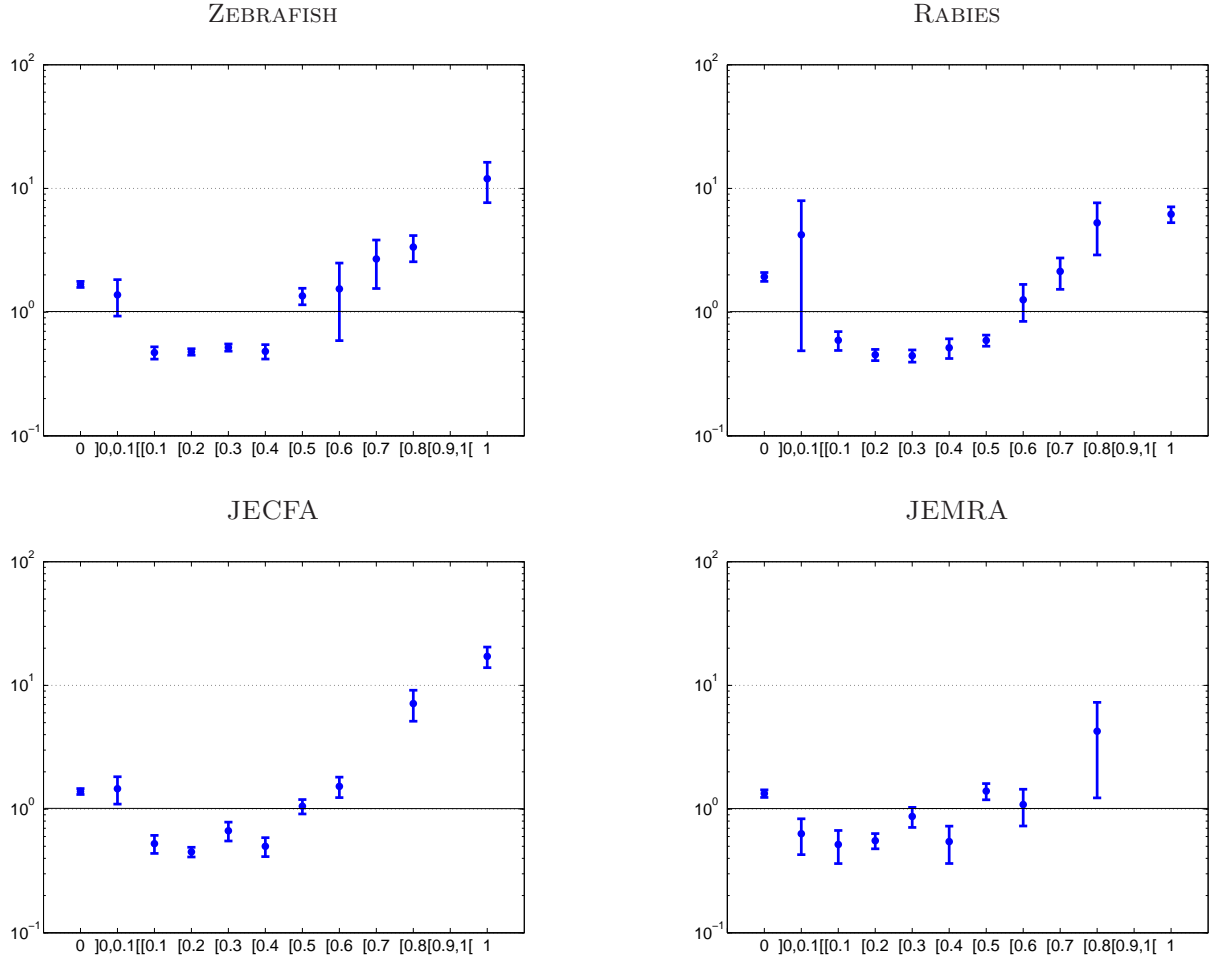


Figure 3: Propensity for **proportions of experts per article**, from our real data vs expected from our random theoretical model — averaged over all years, then over all concepts. (Error bars correspond to 95% confidence intervals with respect to concept averages.)

of neophytes, are also significantly more likely, suggesting that part of the use of new concepts is also due to teams almost completely new to such concepts (even if, as is proved by (H1), these very teams are still more likely to stem from repeated collaborations). Put bluntly, new concept usage, and thus part of innovation, appears to stem both from teams significantly ignorant of such concepts and from teams globally knowledgeable about such concepts.

From this observation that “all-experts” and “all-neophytes” teams are more likely, we may expect that such teams stem from underlying groups (either still working on the same topic, or working on a new topic, respectively) and thus have a higher social hypergraphic repetition ratio. Similarly, those teams stemming from underlying groups are likely to carry normal, specialized science and have higher semantic hypergraphic repetition ratio (or lower originality). Fig-

ure 6 sheds light on these issues by comparing average hypergraphic repetition ratios with expertise ratios. In particular, we observe that teams with a balanced composition of experts have a higher social originality (lower social hypergraphic repetition ratio), yet semantic originality remains constant across various values of expertise ratios. This partially confirms **(HII)** as regards social originality and partially invalidates it as regards semantic originality: indeed, social originality is increased when there is a mixed proportion of experts, but not semantic originality.

5 Concluding remarks

We presented a formal framework to appraise the underpinnings of collaboration formation with a hypergraphic approach which encompasses both the meso-

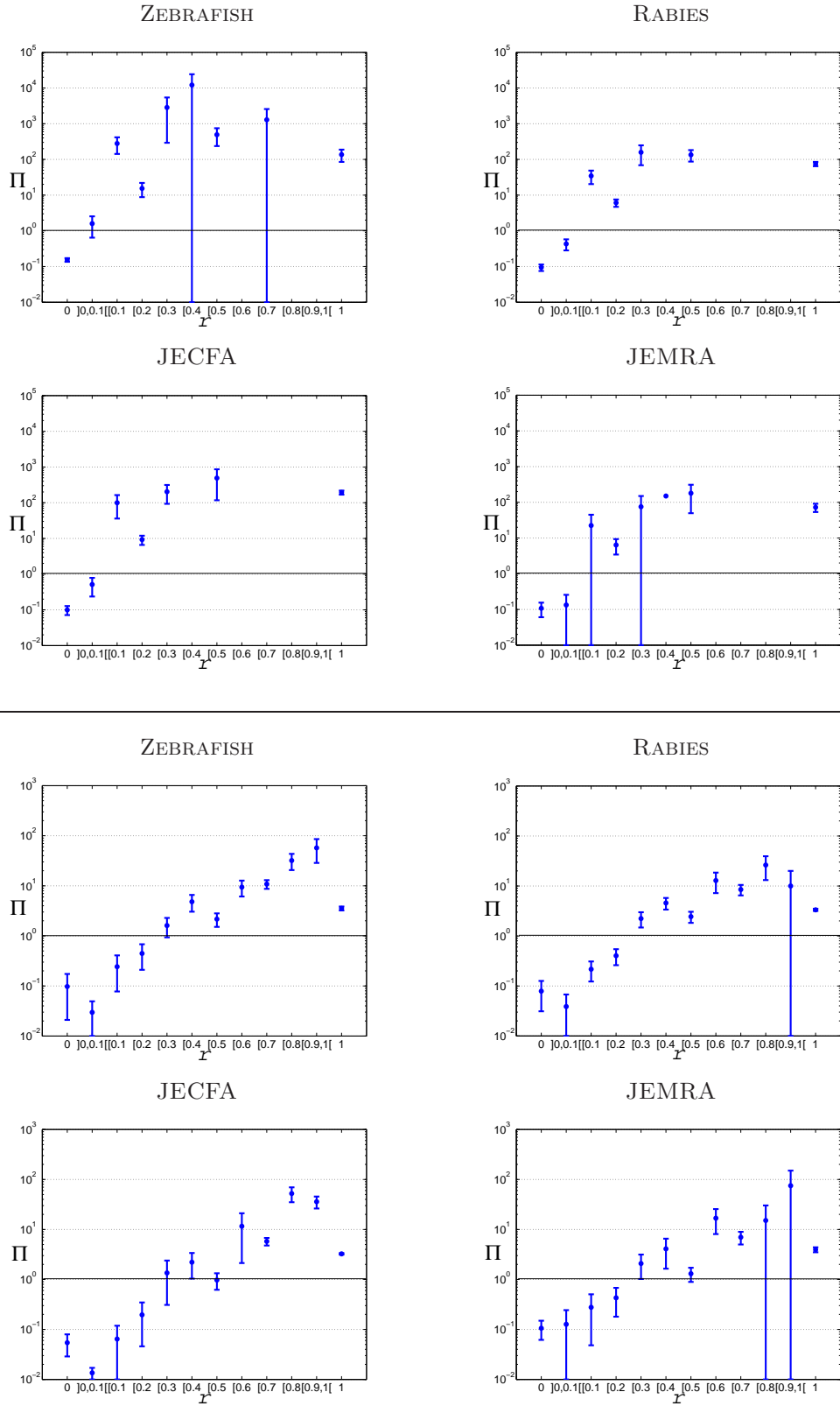


Figure 4: Propensity of team formation (random hypergraph vs. real data) with respect to hypergraphic repetition ratios for **agents** (*top*) and **concepts** (*bottom*). (Values are averaged over all years, error bars correspond to 95% confidence intervals with respect to these averages.)

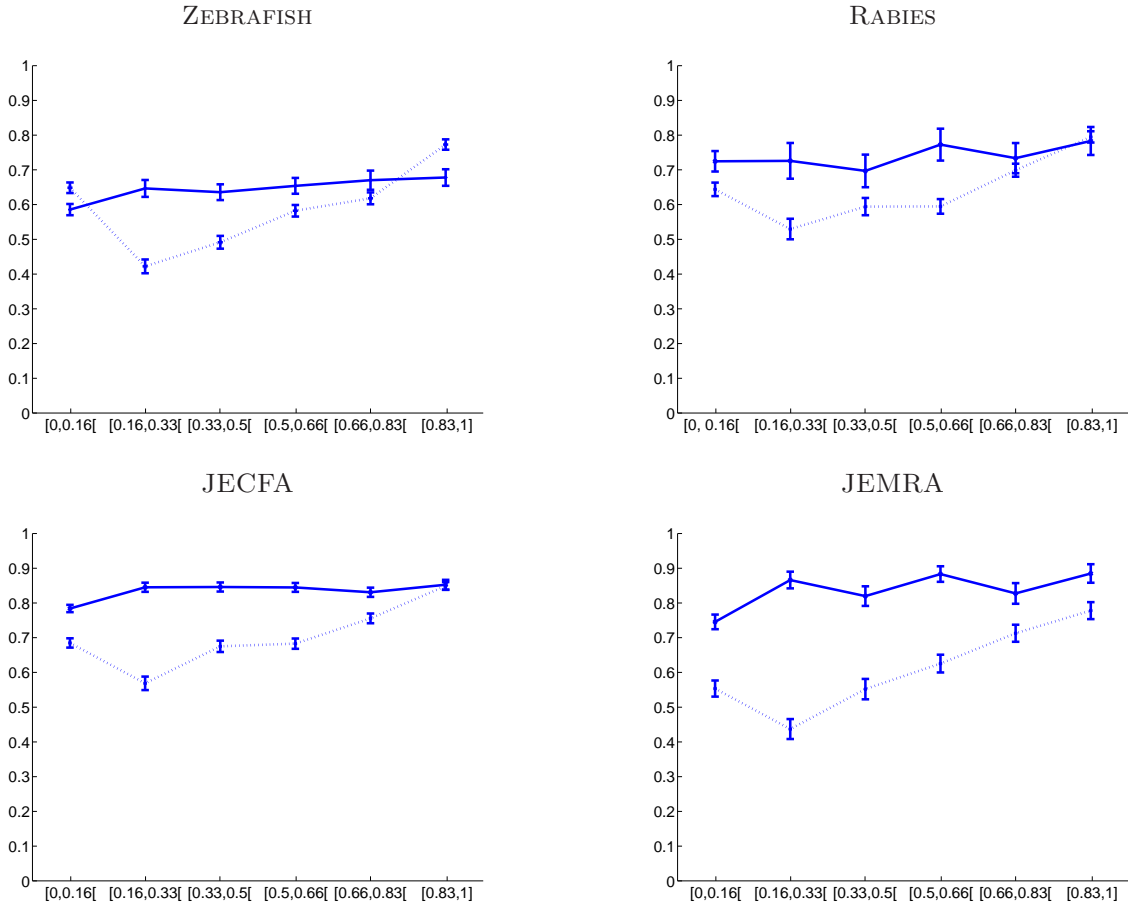


Figure 6: Average hypergraphic repetition ratios (y -axis) with respect to expertise ratios (x -axis): social (*dashed line*) and semantic (*plain line*) cases. (Error bars correspond to 95% confidence intervals with respect to averages on each expertise ratio bin (in abscissa), such as e.g. $[0, 0.1[.$)

level of teams and the joint dynamics of social and semantic features. This allowed the quantitative estimation of the relative strength of social and semantic patterns behind academic team formation, by empirically studying several communities of scientists and estimating how the composition of teams, both cognitively and socially, diverges from a null hypothesis where collaborators and/or topics would be randomly chosen.

We could thereby confirm several hypotheses as well as invalidate some hypotheses which had been established in a relatively qualitative fashion in the literature, or in a possibly misleading *dyadic* form. More precisely, our measurements suggest a mechanism of team formation based on (i) a high likeliness to repeat previous collaborations patterns, not only dyadic but also n -adic interactions ($n \geq 3$) and (ii) a sensible confinement of groups of individuals, whose collaborations appear to depend largely on the history of team memberships, and, similarly, a sensible seman-

tic confinement where associations of concepts depend largely on the repetition of previous associations. On the whole however, the originality of a paper does not seem to stem from an original composition of the underlying team, while a polarization appears between groups made of experts only or made of non-experts only, which altogether correspond to collectives exhibiting a high rate of repeated interactions.

Perspectives on models of academic collaboration. Taking into account an implicit group structure, both at a social and at a socio-semantic level, as evidenced by the data, is likely to faithfully account for the structure of academic collaboration networks. Indeed, the underlying low-level dynamics is plausibly closer to hypergraphic team formation mechanisms than would be allowed by a design based on dyadic interactions only. As said before, this should not yield a lack of organizational thinking regarding the underpinnings of scientific production: beyond the step that constitutes our present contribution, an exhaustive ap-

proach about this type of collaboration mechanisms would indeed have to involve both epistemic hypergraphs and organizational features. In this respect, while we claim and show that hypergraphs make it possible to capture some interesting processes of team-based, knowledge-intensive production systems, we also emphasize that the richness of organizational mechanisms should not be shadowed by this formalism.

In line with our results, it should also be possible to determine which features, at the level-team, favor better collaborations — not only in terms of semantic originality, but also in terms of quality and creativity of output, in a broad sense.

Acknowledgements. This work was partially supported by the Future and Emerging Technologies programme FP7-COSI-ICT of the European Commission through project QLectives (grant no.: 231200). We thank David Chavalarias and several anonymous reviewers for their useful comments.

References

- Adams, J. D., Black, G. C., Clemmons, J. R., Stephan, P. E. (2005), Scientific teams and institutional collaborations: Evidence from U.S. universities, 1981–1999, *Research Policy*, 34:259–285.
- Ancona, D. G., Caldwell, D. F. (1992), Demography and design: predictors of new product team performance, *Organization Science*, 3:321–341.
- Barabási, A.-L., Jeong, H., Ravasz, R., Neda, Z., Vicsek, T., Schubert, T. (2002), Evolution of the social network of scientific collaborations, *Physica A*, 311:590–614.
- Bollen, K. A., Hoyle, R. H. (1990), Perceived cohesion: A conceptual and empirical examination, *Social Forces*, 69:479–504.
- Breiger, R. L. (1974), The duality of persons and groups, *Social Forces*, 53:181–190.
- Breiger, R. L. (1990), Social control and social networks: a model from Georg Simmel, in: C. Calhoun, M. Meyer, W. R. Scott (Eds.) *Structures of Power and Constraint: Papers in Honor of Peter M. Blau*, Cambridge University Press, pp. 453–476.
- Bryant, S. L., Forte, A., Bruckman, A. (2005), Becoming wikipedia: Transformation of participation in a collaborative online encyclopedia, in: *Proc. of Group’05, Sanibel Island, FL, USA*.
- Callon, M. (1986), Some elements of a sociology of translation: domestication of the scallops and the fishermen of StBrieuc Bay, *Power, Action and Belief: A New Sociology of Knowledge*, 32:196–233.
- Callon, M. (1994), Is science a public good?, *Science, Technology & Human Values*, 19:395–424.
- Callon, M., Law, J., Rip, A. (1986), *Mapping the dynamics of science and technology*, MacMillan Press, London.
- Chubin, D. E. (1976), The conceptualization of scientific specialties, *The Sociological Quarterly*, 17:448–476.
- Constant, D., Sproull, L., Kiesler, S. (1996), The kindness of strangers: The usefulness of electronic weak ties for technical advice, *Organization Science*, 7:119–135.
- Cowan, R., Jonard, N., Zimmermann, J.-B. (2002), The joint dynamics of networks and knowledge, *Computing in Economics and Finance* 354, Society for Computational Economics.
- Crane, D. (1969), Social structure in a group of scientists: a test of the ‘invisible college’ hypothesis, *American Sociological Review*, 34:335–352.
- Davis, G. F., Greve, H. R. (1996), Corporate elite networks and governance changes in the 1980s, *American Journal of Sociology*, 103:1–37.
- deB. Beaver, D. (1986), Collaboration and teamwork in physics, *Czech Journal of Physics B*, 36:14–18.
- deB. Beaver, D., Rosen, R. (1978–1979), Studies in scientific collaboration. Parts I, II, III., *Scientometrics*, 1:65–84, 133–149, 231–245.
- Faulkner, R. R., Anderson, A. B. (1987), Short-term projects and emergent careers: Evidence from hollywood, *American Journal of Sociology*, 92:879–909.
- Freeman, L. C. (2003), Finding social groups: A meta-analysis of the Southern women data, in: R. Breiger, K. Carley, P. Pattison (Eds.) *Dynamic Social Network Modeling and Analysis*, The National Academies Press, Washington, D.C., pp. 39–97.
- Friedkin, N. E. (2004), Social cohesion, *Annual Review of Sociology*, 30:409–425.
- Guimera, R., Uzzi, B., Spiro, J., Amaral, L. A. N. (2005), Team assembly mechanisms determine collaboration network structure and team performance, *Science*, 308:697–702.
- Haas, P. (1992), Introduction: epistemic communities and international policy coordination, *International Organization*, 46:1–35.
- Jones, B. F., Wuchty, S., Uzzi, B. (2008), Multi-university research teams: Shifting impact, geography, and stratification in science, *Science*, 322:1259–1262.
- Jones, C., Hesterly, W. S., Borgatti, S. P. (1997), A general theory of network governance: Exchange conditions and social mechanisms, *Academy of Management Review*, 22:911–945.

- Karsai, I., Penzes, Z. (1993), Comb building in social wasps: Self-organization and stigmergic script, *Journal of Theoretical Biology*, 161:505–525.
- Katz, J. S., Martin, B. R. (1997), What is research collaboration?, *Research Policy*, 26:1–18.
- Kogut, B., Metiu, A. (2001), Open-source software development and distributed innovation, *Oxford Review of Economic Policy*, 17:248–264.
- Laband, D. N., Tollison, R. D. (2000), Intellectual collaboration, *The Journal of Political Economy*, 108:632–662.
- Larédo, P. (1995), Structural effects of ec rt & d programmes, *Scientometrics*, 34:473–487.
- Larédo, P. (1998), The networks promoted by the framework programme and the questions they raise about its formulation and implementation, *Research Policy*, 27:589–598.
- Latour, B., Woolgar, S. (1979), *Laboratory Life: The Social Construction of Scientific Facts*, Sage Publications, Beverly Hills.
- Lazega, E., Jourda, M.-T., Mounier, L., Stofer, R. (2008), Catching up with big fish in the big pond? multi-level network analysis through linked design, *Social Networks*, 30:159–176.
- Leahey, E., Reikowsky, R. C. (2008), Research specialization and collaboration patterns in sociology, *Social Studies of Science*, 38:425–440.
- Levrel, J. (2006), Wikipédia, un dispositif médiatique de publics participants, *Réseaux*, 24:185–218.
- Lott, A. J., Lott, B. E. (1965), Group cohesiveness as interpersonal attraction: a review of relationships with antecedent and consequent variables, *Psychological Bulletin*, 64:259–309.
- McPherson, M., Smith-Lovin, L. (2002), Cohesion and membership duration: linking groups, relations and individuals in an ecology of affiliation, *Advances in Group Processes*, 19:1–36.
- McPherson, M., Smith-Lovin, L., Cook, J. M. (2001), Birds of a feather: Homophily in social networks, *Annual Review of Sociology*, 27:415–444.
- Melin, G., Persson, O. (1996), Studying research collaboration usign co-authorships, *Scientometrics*, 36:363–377.
- Miles, R. E., Snow, C. C. (1996), Organizations: New concepts for new forms. a reader in industrial organization, in: P. J. Buckley, J. Michie (Eds.) *Firms, Organizations and Contracts*, Oxford: Oxford University Press, pp. 429–441.
- Moody, J. (2004), The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999, *American Sociological Review*, 69:213–238.
- Mullins, N. C. (1972), The development of a scientific specialty: The phage group and the origins of molecular biology, *Minerva*, 10:51–82.
- Newman, M. E. J. (2001), Scientific collaboration networks. I. Network construction and fundamental results, and II. Shortest paths, weighted networks, and centrality, *Physical Review E*, 64:016131 & 016132.
- Newman, M. E. J., Strogatz, S., Watts, D. (2001), Random graphs with arbitrary degree distributions and their applications, *Physical Review E*, 64:026118.
- Noyons, E. C. M., van Raan, A. F. J. (1998), Monitoring scientific developments from a dynamic perspective: self-organized structuring to map neural network research, *Journal of the American Society for Information Science*, 49:68–81.
- Powell, W. W. (1990), Neither market nor hierarchy: Network forms of organization, *Research in Organizational Behavior*, 12:295–336.
- Ramasco, J. J., Dorogovtsev, S. N., Pastor-Satorras, R. (2004), Self-organization of collaboration networks, *Physical Review E*, 70:036106.
- Rodriguez, M. A., Pepe, A. (2008), On the relationship between the structural and socioacademic communities of a coauthorship network, *Journal of Informetrics*, 2:195–201.
- Roth, C. (2006), Co-evolution in epistemic networks – reconstructing social complex systems, *Structure and Dynamics: eJournal of Anthropological and Related Sciences*, 1(3).
- Roth, C., Cointet, J.-P. (2010), Social and semantic coevolution in knowledge networks, *Social Networks*, 32:16–29.
- Ruef, M. (2002), A structural event approach to the analysis of group composition, *Social Networks*, 24:135–160.
- Ruggie, J. G. (1975), International responses to technology: Concepts and trends, *International Organization*, 29:557–583.
- Simmel, G. (1898), The persistence of social groups, *American Journal of Sociology*, 3:662.
- Smangs, M. (2006), The nature of the business group: A social network perspective, *Organization*, 13:889–909.
- Stokols, D., Hall, K. L., Taylor, B. K., Moser, R. P. (2008), The science of team science, *American Journal of Preventive Medicine*, 35:S78–S89.
- Uzzi, B., Spiro, J. (2005), Collaboration and creativity: the small-world problem, *American Journal of Sociology*, 111:447–504.
- Wagner, C. S., Leydesdorff, L. (2005), Network structure, self-organization, and the growth of international collaboration in science, *Research Policy*, 34:1608–1618.

Welser, H. T., Gleave, E., Fischer, D., Smith, M. (2007), Visualizing the signatures of social roles in online discussion groups, *Journal of Social Structure*, 8.

Wuchty, S., Jones, B., Uzzi, B. (2007), The increasing dominance of teams in the production of knowledge, *Science*, 316:1036–1039.

A Weighting functions

A weighted hypergraphic repetition rate could be written as follows:

$$r_t(\mathfrak{e}) = \frac{\sum_{\substack{\mathfrak{e}' \subseteq \mathfrak{e} \\ |\mathfrak{e}'| \geq 2}} w_{\mathfrak{e}}(|\mathfrak{e}'|) \cdot \rho_t(\mathfrak{e}')}{\sum_{i \in \{2, \dots, |\mathfrak{e}|\}} w_{\mathfrak{e}}(i) \binom{|\mathfrak{e}|}{i}}$$

where $w_{\mathfrak{e}}$ is a weight function (given \mathfrak{e} , $w_{\mathfrak{e}} : \mathbb{N} \rightarrow \mathbb{R}$) which makes it possible to give more or less weight to particular subset sizes.

For instance:

- taking $w_{\mathfrak{e}}(i) = 1$, i.e. actually no weighting as has been used in the paper,

$$r_t(\mathfrak{e}) = \frac{1}{2^{|\mathfrak{e}|} - |\mathfrak{e}| - 1} \sum_{\substack{\mathfrak{e}' \subseteq \mathfrak{e} \\ |\mathfrak{e}'| \geq 2}} \rho_t(\mathfrak{e}')$$

- if instead $w_{\mathfrak{e}}(i) = i$, i.e. weighting proportional to the size of the considered subset,

$$r_t(\mathfrak{e}) = \frac{1}{|\mathfrak{e}|(2^{|\mathfrak{e}|-1} - 1)} \sum_{\substack{\mathfrak{e}' \subseteq \mathfrak{e} \\ |\mathfrak{e}'| \geq 2}} |\mathfrak{e}'| \rho_t(\mathfrak{e}')$$

- if finally $w_{\mathfrak{e}}(i) = \binom{|\mathfrak{e}|}{i}^{-1}$, i.e. weighting proportional to the number of possible subsets of size $|\mathfrak{e}|$ in a set of size i ,

$$r_t(\mathfrak{e}) = \sum_{\substack{\mathfrak{e}' \subseteq \mathfrak{e} \\ |\mathfrak{e}'| \geq 2}} \frac{\rho_t(\mathfrak{e}')}{\binom{|\mathfrak{e}|}{|\mathfrak{e}'|}}$$